

# Time series forecasting: model evaluation and selection using nonparametric risk bounds\*

Daniel J. McDonald<sup>†</sup>, Cosma Rohilla Shalizi<sup>‡</sup>, and Mark Schervish<sup>§</sup>

<sup>†</sup>Department of Statistics, Indiana University Bloomington

<sup>§</sup>Department of Statistics, Carnegie Mellon University

<sup>‡</sup>Santa Fe Institute

Version: December 4, 2012

## Abstract

We derive generalization error bounds — bounds on the expected inaccuracy of the predictions — for traditional time series forecasting models. Our results hold for many standard forecasting tools including autoregressive models, moving average models, and, more generally, linear state-space models. These bounds allow forecasters to select among competing models and to guarantee that with high probability, their chosen model will perform well without making strong assumptions about the data generating process or appealing to asymptotic theory. We motivate our techniques with and apply them to standard economic and financial forecasting tools — a GARCH model for predicting equity volatility and a dynamic stochastic general equilibrium model (DSGE), the standard tool in macroeconomic forecasting. We demonstrate in particular how our techniques can aid forecasters and policy makers in choosing models which behave well under uncertainty and mis-specification.

**Keywords:** Generalization error, Prediction risk, Model selection.

## 1 Introduction

Generalization error bounds are probabilistically valid, non-asymptotic tools for characterizing the predictive ability of forecasting models. This methodology is fundamentally about choosing particular prediction functions out of some class of plausible alternatives so that, with high reliability, the resulting predictions will be nearly as accurate as possible (“probably approximately correct”). While many of these results are useful only for classification problems (i.e., predicting binary variables) and for independent and identically distributed (IID) data, this paper adapts and extends these methods to time series models, so that economic and financial forecasting techniques can be evaluated rigorously. In particular, these methods control the expected accuracy of future predictions from mis-specified models based on finite samples. This allows for immediate model comparisons which neither appeal to asymptotics nor make strong assumptions about the data-generating process, in stark contrast to such popular model-selection tools as AIC.

To fix ideas, imagine IID<sup>1</sup> data  $((Y_1, X_1), \dots, (Y_n, X_n))$  with  $(Y_i, X_i) \in \mathcal{Y} \times \mathcal{X}$ , some prediction function  $f : \mathcal{X} \rightarrow \mathcal{Y}$ , and a loss function  $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$  which measures the cost of bad predictions. The *generalization error* or *risk* of  $f$  is

$$R(f) := \mathbb{E}[\ell(Y, f(X))] \quad (1)$$

where the expectation is taken with respect to  $\mathbb{P}$ , the joint distribution of  $(Y, X)$ . The generalization error measures the inaccuracy of our predictions when we use  $f$  on future data, making it a natural criterion for

\*Email: [dajmcdon@indiana.edu](mailto:dajmcdon@indiana.edu), [cshalizi@cmu.edu](mailto:cshalizi@cmu.edu), [mark@cmu.edu](mailto:mark@cmu.edu). This work is partially supported by a grant from the Institute for New Economic Thinking. CRS was also partially supported by NIH Grant # 2 R01 NS047493. The authors wish to thank David N. Dejong, Larry Wasserman, Alessandro Rinaldo and Darren Homrighausen for valuable suggestions.

<sup>1</sup>The IID assumption here is just for ease of exposition; we develop dependent-data results at length below.

model selection, and a target for performance guarantees. To actually calculate the risk, we would need to know the data-generating distribution  $\mathbb{P}$  and have a single fixed prediction function  $f$ , neither of which is common. Because explicitly calculating the risk is infeasible, forecasters typically try to estimate it, which calls for detailed assumptions on  $\mathbb{P}$ . The alternative we employ here is to find upper bounds on risk which hold uniformly over large classes of models  $\mathcal{F}$  from which some particular  $f$  is chosen, possibly in a data dependent way, and uniformly over distributions  $\mathbb{P}$ .

Our main results in [Section 4](#) assert that for wide classes of time series models (including VARs and state-space models), the expected cost of poor predictions is bounded by the model’s in-sample performance inflated by a term which balances the amount of observed data with the complexity of the model. The bound holds with high probability under the unknown distribution  $\mathbb{P}$  assuming only mild conditions — existence of some moments, stationarity, and the decay of temporal dependence as data points become widely separated in time. As a preview, the following provides the general form of the result. Specific results which have this flavor are [Theorem 4.3](#) and [Theorem 4.6](#) and their corollaries. We give applications in [Section 5](#).

**Result.** *Given a time series  $Y_1, \dots, Y_n$  satisfying some mild conditions and a prediction function  $f$  chosen from a class of functions  $\mathcal{F}$  (possibly by using the observed sample), then, with probability at least  $1 - \eta$ ,*

$$R(f) \leq \hat{R}_n(f) + C_{\mathcal{F}}(\eta, n) \quad (2)$$

*where  $R(f)$  is the expected cost of making prediction errors on new samples,  $\hat{R}_n(f)$  is the average cost of in-sample prediction errors,  $C_{\mathcal{F}}(\eta, n) \geq 0$  balances the complexity of the model from which  $f$  was chosen with the amount of data used to choose it.*

There are many ways to estimate the generalization error, and a comprehensive review is beyond the scope of this paper. Traditionally, time series analysts have performed model selection by a combination of empirical risk minimization, more-or-less quantitative inspection of the residuals, and penalties like AIC. In many applications, however, what really matters is prediction, and none of these techniques work to control generalization error, especially for mis-specified models. Empirical cross-validation is a partial exception, but it is tricky for time series; see Racine [\[44\]](#) and references therein. In economics, forecasters have long recognized the difficulties with these methods, preferring to use a pseudo-cross validation approach instead: choose a prediction function using the initial portion of a data set and evaluate its performance on the remainder (c.f. [\[2, 16, 19, 50\]](#)). This procedure provides approximate solutions to the problem of estimating the generalization error, but it can be biased toward overfitting — giving too much credence to the observed data — and hence tends to underestimate the true risk for at least three reasons. First, the held-out data, or test set, is used to evaluate the performance of competing models despite the fact that it was already partially used to build those models. For instance, the recent housing and financial crises have precipitated attempts to enrich existing models with mechanisms designed to enhance their ability to predict just such a crisis (c.f. [\[21–23\]](#)). Second, the test set may reflect only a small sampling of possible phenomena which could occur. Finally, large departures from the normal course of events such as the recessions in 1980–82 and periods before 1960 are often ignored, as in [\[19\]](#). While these periods are considered rare and perhaps unpredictable, models which are robust to these sorts of disruptive events will lead to more accurate predictions in future times of turmoil.

In contrast to the model evaluation techniques typically employed in the literature, generalization error bounds provide rigorous control over the predictive risk as well as reliable methods of model selection. They are robust to wide classes of data generating processes and are finite-sample rather than asymptotic in nature. In a broad sense, these methods give confidence intervals which are constructed based on concentration of measure results rather than appeals to asymptotic normality. The results are easy to understand and can be reported to policy makers interested in the quality of the forecasts. Finally, the results are agnostic about the model’s specification: it does not matter if the model is wrong, whether the parameters have interpretable economic meaning, or whether the estimation of the parameters is performed only approximately (linearized DSGEs or MCMC). In all of these cases, we can still make strong claims about the ability of the model to predict the future.

The bounds we derive here are the first of their kind for the time series models typically used in applied settings — finance, economics, engineering, etc. — but there are results for other models more common to computer science (cf. Meir [\[37\]](#), Mohri and Rostamizadeh [\[38, 39\]](#)). Those results require bounded loss

functions, making them less general than ours, as well as hinging on specific forms of regularization which are rarely used in time series. Furthermore, they rely on prediction functions  $f : \mathcal{X} \rightarrow \mathcal{Y}$  where the dependence occurs in the  $\mathcal{X}$  space. Therefore, these results are extensible to AR models or others which depend on only the most recent past (assuming appropriate model space constraints are satisfied) but not, for instance, to standard state-space models. For another view on this problem, [36] shows that stationarity alone can be used to regularize an AR model following the results in [38], but leads to bounds which are much worse than those given here, despite the stricter assumption of bounded loss.

The meaning of such results for forecasters, or for those whose scientific aims center around prediction of empirical phenomena, is plain: they provide objective ways of assessing how good their models really are. There are, of course, other uses for scientific models: for explanation, for the evaluation of counterfactuals (especially, in economics, comparing the consequences of different policies), and for welfare calculations. Even in those cases, however, one must ask *why this model rather than another?*, and the usual answer is that the favored model approximates reality better than the alternative — it gets the structure approximately right. Empirical evidence for structural correctness, in turn, usually takes the form of an argument from empirical success: *it would be very surprising if this model fit the data so well when it got the structure wrong* [33]. Our results, which directly address the inference from past data-matching to future performance, are thus relevant even to those who do not aim at prediction as such.

The remainder of this paper is structured as follows. Section 2 provides motivation and background for our results, giving intuition in the IID setting by focusing on concentration of measure ideas and characterizations of model complexity. Section 3 gives the explicit assumptions we make and describes how to leverage powerful ideas from time series to generalize the IID methods. Section 4 states and proves risk bounds for the time series forecasting setting, while we demonstrate how to use the results in Section 5 and give some properties of those results in Section 6. Finally, Section 7 concludes and illustrates the path toward generalizing our methods to more elaborate model classes.

## 2 Statistical learning theory

Our goal is to control the risk of predictive models, i.e., their expected inaccuracy on new data from the same source as that used to fit the model. To orient readers new to this approach, we sketch how classical results in the IID setting are obtained.

Let  $f : \mathcal{X} \rightarrow \mathcal{Y}$  be some function used for making predictions of  $Y$  from  $X$ . We define a loss function  $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$  which measures the cost of making poor predictions. Throughout this paper, we will assume that  $\ell(y, y')$  is a function solely of the difference  $y - y'$  where  $\ell(\cdot)$  is nonnegative and  $\ell(0) = 0$ . For the remainder of the paper, we take the liberty of denoting that function  $\ell(y - y')$ . Then the risk of any predictor  $f \in \mathcal{F}$  (where  $f$  is fixed independently of the data) is given by

$$R(f) = \mathbb{E}[\ell(Y - f(X))], \quad (3)$$

where  $(X, Y) \sim \mathbb{P}$ . The risk or generalization error is the expected cost of using  $f$  to predict  $Y$  from  $X$  on a new observation.

Since the true distribution  $\mathbb{P}$  is unknown, so is  $R(f)$ , but we can try to estimate it based on our observed data. The *training error* or *empirical risk* of  $f$  is

$$\hat{R}_n(f) := \frac{1}{n} \sum_{i=1}^n \ell(Y_i - f(X_i)). \quad (4)$$

In other words, the in-sample training error,  $\hat{R}_n(f)$ , is the average loss over the actual training points. Because the true risk is an expectation value, we can say that

$$\hat{R}_n(f) = R(f) + \gamma_n(f), \quad (5)$$

where  $\gamma_n(f)$  is a mean-zero noise variable that reflects how far the training sample departs from being perfectly representative of the data-generating distribution. By the law of large numbers, for each fixed  $f$ ,  $\gamma_n(f) \rightarrow 0$  as  $n \rightarrow \infty$ , so, with enough data, we have a good idea of how well any given function will generalize to new data.

However, forecasters rarely have the luxury of a theory which fixes for them, in advance of the data, a single function  $f$ , free of adjustable parameters. Rather, there is a class of plausible functions  $\mathcal{F}$ , possibly indexed by some parameters  $\theta \in \Theta$ , which we will call “a model”. One picks out a single function (chooses one particular parameter point) from the model via some method — maximum likelihood, Bayesian updating, indirect inference, ad hoc methods — which often amounts to minimizing the in-sample loss. In this case, the result is

$$\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \hat{R}_n(f) = \operatorname{argmin}_{f \in \mathcal{F}} (R(f) + \gamma_n(f)). \quad (6)$$

Tuning the parameters so that  $\hat{f}$  fits the training data well thus conflates predicting future data well (low true risk  $R(\hat{f})$ ) with exploiting the accidents and noise of the training data (large negative finite-sample noise  $\gamma_n(\hat{f})$ ). The true risk of  $\hat{f}$  will generally be bigger than its in-sample risk precisely because we picked it to match the data well. In doing so,  $\hat{f}$  ends up reproducing some of the noise in the data and therefore will not generalize as well as  $\hat{R}_n(\hat{f})$  suggests. The difference between the true and apparent risk depends on the magnitude of the sampling fluctuations:

$$R(\hat{f}) - \hat{R}_n(\hat{f}) \leq \sup_{f \in \mathcal{F}} |\gamma_n(f)| = \Gamma_n(\mathcal{F}). \quad (7)$$

The main goal of statistical learning theory is to mathematically control  $\Gamma_n(\mathcal{F})$ , finding tight bounds on this quantity which make weak assumptions about the unknown data-generating process; i.e., to bound over-fitting. Using more flexible models (allowing more general functional forms or distributions, adding parameters, etc.) has two contrasting effects. On the one hand, it improves the best possible accuracy, lowering the minimum of the true risk. On the other hand, it increases the ability to, as it were, memorize noise for any fixed sample size  $n$ . This qualitative observation — a form of the bias-variance trade-off from basic estimation theory — can be made usefully precise by quantifying the complexity of model classes. A typical result is a confidence bound on  $\Gamma_n$  (and hence on the over-fitting), which says that with probability at least  $1 - \eta$ ,

$$\Gamma_n(\mathcal{F}) \leq \Phi(\Psi(\mathcal{F}), n, \eta), \quad (8)$$

where  $\Psi(\cdot)$  measures the complexity of the model  $\mathcal{F}$ .

To give specific forms of  $\Phi(\cdot)$ , we need to show that, for a particular  $f$ ,  $R(f)$  and  $\hat{R}_n(f)$  will be close to each other for each fixed  $n$ , without knowledge of the distribution of the data. We also need to understand the complexity,  $\Psi(\mathcal{F})$ , so that we can claim  $R(f)$  and  $\hat{R}_n(f)$  will be close *uniformly* over all  $f \in \mathcal{F}$ . Together these two pieces tell us, despite little knowledge of the data generating process, how bad the  $\hat{f}$  which we choose will be at forecasting future observations.

## 2.1 Concentration

The first step to controlling the difference between the empirical and expected risk is to show that for each  $f \in \mathcal{F}$ ,  $R(f) - \hat{R}_n(f)$  is small with high probability. The following is a standard result (c.f. [55] or [12]).

**Theorem 2.1.** *Suppose that  $0 \leq \ell(y, y') \leq K < \infty$ . Then for each  $f \in \mathcal{F}$ ,*

$$\mathbb{P} \left( \left| R(f) - \hat{R}_n(f) \right| \geq \epsilon \right) \leq 2 \exp \left\{ -\frac{2n\epsilon^2}{K^2} \right\}. \quad (9)$$

*Proof.* The proof begins by using an exponential version of Markov’s inequality. For a fixed  $f$ , we have  $\mathbb{E} [\hat{R}_n(f)] = R(f)$ . Therefore

$$\mathbb{P} \left( R(f) - \hat{R}_n(f) > \epsilon \right) = \mathbb{P} \left( \exp\{s(R(f) - \hat{R}_n(f))\} \geq \exp\{s\epsilon\} \right) \quad (10)$$

$$\leq \frac{\mathbb{E} \left[ \exp\{s(R(f) - \hat{R}_n(f))\} \right]}{\exp\{s\epsilon\}}. \quad (11)$$

We can bound the moment generating function,  $\mathbb{E} \left[ \exp\{s(R(f) - \widehat{R}_n(f))\} \right]$  via Hoeffding's inequality [26]:

$$\mathbb{E}[\exp\{s(R(f) - \widehat{R}_n(f))\}] = \prod_{i=1}^n \mathbb{E} \left[ \exp \left\{ \frac{s}{n} [R(f) - \ell(Y_i, f(X_i))] \right\} \right] \quad (12)$$

$$\leq \prod_{i=1}^n \exp \left\{ \frac{s^2 K^2}{8n^2} \right\} = \exp \left\{ \frac{s^2 K^2}{8n} \right\}. \quad (13)$$

With this result, we have

$$\mathbb{P} \left( R(f) - \widehat{R}_n(f) > \epsilon \right) \leq \exp\{-s\epsilon\} \exp \left\{ \frac{s^2 K^2}{8n} \right\}. \quad (14)$$

This holds for all  $s > 0$ , so we can minimize the right hand side in  $s$  (this is known as Chernoff's method). The minimum occurs for  $s = 4n\epsilon/K^2$ . Substitution gives

$$\mathbb{P} \left( R(f) - \widehat{R}_n(f) > \epsilon \right) \leq \exp \left\{ -\frac{2n\epsilon^2}{K^2} \right\}. \quad (15)$$

Exactly the same argument holds for  $\mathbb{P}(R(f) - \widehat{R}_n(f) < -\epsilon)$ , so by a union bound, we have the result.  $\blacksquare$

This result is quite powerful: it says that the probability of observing data which will result in a training error much different from the expected risk goes to zero exponentially with the size of training set. The only assumption necessary was that  $\ell(y - y') \leq K$ . In fact, even this assumption can be removed and replaced with some moment assumptions, as will be done for our main results below.

**Theorem 2.1** holds for the single function  $f$ , and we want a similar result to hold uniformly over all functions  $f \in \mathcal{F}$  and in particular, any  $\widehat{f}$  that we might choose using the training data, i.e., we wish to bound  $\mathbb{P} \left( \sup_{f \in \mathcal{F}} |R(f) - \widehat{R}_n(f)| > \epsilon \right)$ . How can we achieve this extension?

## 2.2 Capacity

For “small” models, we can just count the number of functions in the class and take the union bound. Suppose that  $\mathcal{F} = \{f_1, \dots, f_N\}$ . Then we have

$$\mathbb{P} \left( \sup_{1 \leq i \leq N} |R(f_i) - \widehat{R}_n(f_i)| > \epsilon \right) \leq \sum_{i=1}^N \mathbb{P} \left( |R(f_i) - \widehat{R}_n(f_i)| > \epsilon \right) \quad (16)$$

$$\leq N \exp \left\{ -\frac{2n\epsilon^2}{K^2} \right\}, \quad (17)$$

by Theorem 2.1. Most interesting models are not small in this sense, but similar results hold when model size is measured appropriately.

There are a number of measures for the size or capacity of a model. Algorithmic stability [4, 5, 28] quantifies the sensitivity of the chosen function to small perturbations to the data. Similarly, maximal discrepancy [53] asks how different the predictions could be if two functions are chosen using two separate data sets. A more direct, functional-analytic approach partitions  $\mathcal{F}$  into equivalence classes under some metric, leading to covering numbers [42, 43]. Rademacher complexity [3] directly describes a model's ability to fit random noise. We focus on a measure which is both intuitive and powerful: Vapnik-Chervonenkis (VC) dimension [52, 53].

VC dimension starts as an idea about collections of sets.

**Definition 2.2.** Let  $\mathbb{U}$  be some (infinite) set and  $S$  a finite subset of  $\mathbb{U}$ . Let  $\mathcal{C}$  be a family of subsets of  $\mathbb{U}$ . We say that  $\mathcal{C}$  shatters  $S$  if for every  $S' \subseteq S$ ,  $\exists C \in \mathcal{C}$  such that  $S' = S \cap C$ .

Essentially,  $\mathcal{C}$  can shatter a set  $S$  if it can pick out every subset of points in  $S$ . This says that the collection  $\mathcal{C}$  is very complicated or flexible. The cardinality of the largest set  $S$  that can be shattered by  $\mathcal{C}$  is the latter's VC dimension.

**Definition 2.3** (VC dimension). *The Vapnik-Chervonenkis (VC) dimension of a collection  $\mathcal{C}$  of subsets of  $\mathbb{U}$  is*

$$\text{VCD}(\mathcal{C}) := \sup\{|S| : S \subseteq \mathbb{U} \text{ and } S \text{ is shattered by } \mathcal{C}\}. \quad (18)$$

To see why this is a “dimension”, we need one more notion.

**Definition 2.4** (Growth function). *The growth function  $G(\mathcal{C}, n)$  of a collection  $\mathcal{C}$  of subsets of  $\mathbb{U}$  is the maximum number of subsets which can be formed by intersecting a set  $S \subset \mathbb{U}$  of cardinality  $n$  with  $\mathcal{C}$ ,*

$$G(n, \mathcal{C}) := \sup_{S \subset \mathbb{U} : |S|=n} |S \wedge \mathcal{C}| \quad (19)$$

The growth function counts how many *effectively* distinct sets the collection contains, when we can only observe what is going on at  $n$  points, not all of  $\mathbb{U}$ . If  $n \leq \text{VCD}(\mathcal{C})$ , then from the definitions  $G(n, \mathcal{C}) = 2^n$ . If the VC dimension is finite, however, and  $n > \text{VCD}(\mathcal{C})$ , then  $G(n, \mathcal{C}) < 2^n$ , and in fact it can be shown [54] that

$$G(n, \mathcal{C}) \leq (n+1)^{\text{VCD}(\mathcal{C})}. \quad (20)$$

This polynomial growth of capacity with  $n$  is why VCD is a “dimension”.

Using VC dimension to measure the capacity of function classes is straightforward. Define the indicator function  $\mathbf{1}_A(x)$  to take the value 1 if  $x \in A$  and 0 otherwise. Suppose that  $f \in \mathcal{F}$ ,  $f : \mathbb{U} \rightarrow \mathbb{R}$ . Each  $f$  corresponds to the set

$$C_f = \{(u, a) : \mathbf{1}_{(0, \infty)}(f(u) - b) = 1, \quad u \in \mathbb{U}, \quad b \in \mathbb{R}\}, \quad (21)$$

so  $\mathcal{F}$  corresponds to the class  $\mathcal{C}_{\mathcal{F}} := \{C_f : f \in \mathcal{F}\}$ . Essentially, the growth function  $G(n, \text{VCD}(\mathcal{F}))$  counts the effective number of functions in  $\mathcal{F}$ , i.e., how many can be told apart using only  $n$  observations. When  $\text{VCD}(\mathcal{F}) < \infty$ , this number grows only polynomially with  $n$ . This observation lets us control the risk over the entire model, providing one of the pillars of statistical learning theory.

**Theorem 2.5** (Vapnik and Chervonenkis [54]). *Suppose that  $\text{VCD}(\mathcal{F}) < \infty$  and  $0 \leq \ell(y, y') \leq K < \infty$ . Then,*

$$\mathbb{P} \left( \sup_{f \in \mathcal{F}} |R(f) - \hat{R}_n(f)| > \epsilon \right) \leq 4(2n+1)^{\text{VCD}(\mathcal{F})} \exp \left\{ -\frac{n\epsilon^2}{K_1^2} \right\}, \quad (22)$$

where  $K_1$  depends only on  $K$  and not  $n$  or  $\mathcal{F}$ .

The proof of this theorem has a similar flavor to the union bound argument given in (17).

This theorem has as an immediate corollary a bound for the out-of-sample risk. Since  $\sup_{f \in \mathcal{F}}$  is inside the probability statement in (22), it applies to both pre-specified and to data-dependent functions, including any  $\hat{f}$  chosen by fitting a model or minimizing empirical risk.

**Corollary 2.6.** *When Theorem 2.5 applies, for any  $\eta > 0$  and any  $f \in \mathcal{F}$ , with probability at least  $1 - \eta$ ,*

$$R(f) \leq \hat{R}_n(f) + K_1 \sqrt{\frac{\text{VCD}(\mathcal{F}) \log(2n+1) + \log 4/\eta}{n}}. \quad (23)$$

The factor  $K_1$  can be calculated explicitly but is unilluminating and we will not need it. Conceptually, the right-hand side of this inequality resembles standard model selection criteria, like AIC or BIC, with in-sample fit plus a penalty term which goes to zero as  $n \rightarrow \infty$ . Here however, the bound holds with high probability despite lack of knowledge of  $\mathbb{P}$  and it has nothing to do with asymptotic convergence: it holds for each  $n$ . It does however hold *only* with high  $\mathbb{P}$  probability, not always.

VC dimension is well understood for some function classes. For instance, if  $\mathcal{F} = \{\mathbf{x} \mapsto \gamma \cdot \mathbf{x} : \gamma \in \mathbb{R}^p\}$  then  $\text{VCD}(\mathcal{F}) = p + 1$ , i.e. it is the number of free parameters in a linear regression plus 1. VC dimension does not always have such a nice relation to the number of free parameters however; the classic example is the model  $\mathcal{F} = \{x \mapsto \sin(\omega x) : \omega \in \mathbb{R}\}$ , which has only one free parameter, but  $\text{VCD}(\mathcal{F}) = \infty$ .<sup>2</sup> At the

<sup>2</sup>This result follows if we can show that for any positive integer  $J$  and any binary sequence  $(r_1, \dots, r_J)$ , there exists a vector  $(x_1, \dots, x_J)$  such that  $\mathbf{1}_{[0,1]}(\sin(\omega x_i)) = r_i$ . If we choose  $x_i = 2\pi 10^{-i}$ , then one can show that taking  $\omega = \frac{1}{2} \left( \sum_{i=1}^J (1 - r_i) 10^i + 1 \right)$  solves the system of equations.



same time, there are model classes (support vector machines) which may have infinitely many parameters but finite VC dimension [11]. This illustrates a further difference between the statistical learning approach and the usual information criteria, which are based on parameter-counting.

The concentration results in [Theorem 2.5](#) and [Corollary 2.6](#) work well for independent data. The first shows how quickly averages concentrate around their expectations: exponentially fast in the size of the data. The second result generalizes the first from a single function to entire function classes. Both results, as stated, depend critically on the independence of the random variables. For time series, we must be able to handle dependent data. In particular, because time-series data are dependent, the length  $n$  of a sample path  $Y_1, \dots, Y_n$  exaggerates how much information it contains. Knowing the past allows forecasters to predict future data (at least to some degree), so actually observing those future data points gives less information about the underlying process than in the IID case. Thus, while in [Theorem 2.1](#) the probability of large discrepancies between empirical means and their expectations decreases exponentially in  $n$ , in the dependent case, the effective sample size may be much less than  $n$  resulting in looser bounds.

### 3 Time series

In moving from the IID setting to time series forecasting, we need a number of modifications to our initial setup. Rather than observing input/output pairs  $(Y_i, X_i)$ , we observe a single sequence of random variables  $Y_{1:n} := (Y_1, \dots, Y_n)$  where each  $Y_i$  takes values in  $\mathbb{R}^p$ .<sup>3</sup> We are interested in using functions which take past observations as inputs and predict future values of the process. Specifically, given data from time 1 to time  $n$ , we wish to predict time  $n + 1$ .

While we no longer presume IID data, we still need to restrict the sort of dependent process we work with. We first remind the reader of the notion of (strict or strong) stationarity.

**Definition 3.1** (Stationarity). *A random sequence  $Y_\infty$  is stationary when all its finite-dimensional distributions are time-invariant: for all  $t$  and all non-negative integers  $i$  and  $j$ , the random vectors  $Y_{t:t+i}$  and  $Y_{t+j:t+i+j}$  have the same distribution.*

Stationarity does not imply that the random variables  $Y_t$  are independent across time  $t$ , only that the unconditional distribution of  $Y_t$  is constant in time. We limit ourselves not just to stationary processes, but also to ones in which widely-separated observations are asymptotically independent. Without this restriction, convergence of the training error to the expected risk could occur arbitrarily slowly, and finite-sample bounds may not exist.<sup>4</sup> The next definition describes the sort of serial dependence which we entertain.

**Definition 3.2** ( $\beta$ -Mixing). *Consider a stationary random sequence  $Y_\infty$  defined on a probability space  $(\Omega, \Sigma, \mathbb{P}_\infty)$ . Let  $\sigma_{i:j} = \sigma(Y_{i:j})$  be the  $\sigma$ -field of events generated by the appropriate collection of random variables. Let  $\mathbb{P}_0$  be the restriction of  $\mathbb{P}_\infty$  to  $\sigma_{-\infty:0}$ ,  $\mathbb{P}_a$  be the restriction of  $\mathbb{P}_\infty$  to  $\sigma_{a:\infty}$ , and  $\mathbb{P}_{0 \otimes a}$  be the restriction of  $\mathbb{P}_\infty$  to  $\sigma(Y_{-\infty:0}, Y_{a:\infty})$ . The coefficient of absolute regularity, or  $\beta$ -mixing coefficient,  $\beta_a$ , is given by*

$$\beta_a := \|\mathbb{P}_0 \times \mathbb{P}_a - \mathbb{P}_{0 \otimes a}\|_{TV}, \quad (24)$$

where  $\|\cdot\|_{TV}$  is the total variation norm. A stochastic process is absolutely regular, or  $\beta$ -mixing, if  $\beta_a \rightarrow 0$  as  $a \rightarrow \infty$ .

This is only one of many equivalent characterizations of  $\beta$ -mixing (see Bradley [6] for others). This definition makes clear that a process is  $\beta$ -mixing if the joint probability of events which are widely separated in time approaches the product of the individual probabilities, i.e., that  $Y_\infty$  is asymptotically independent. Many common time series models are known to be  $\beta$ -mixing, and the rates of decay are known up to constant factors which are functions of the true parameters of the process. Among the processes for which such results are known are ARMA models [40], GARCH models [7], and certain Markov processes — see Doukhan [17] for an overview. Additionally, functions of  $\beta$ -mixing processes are  $\beta$ -mixing, so if  $\mathbb{P}_\infty$  could be specified by a dynamic factor model or DSGE or VAR, the observed data would satisfy this condition.

<sup>3</sup>We can easily generalize this to arbitrary measurable spaces.

<sup>4</sup>In fact, Adams and Nobel [1] demonstrate that for ergodic processes, finite VC dimension is enough to give consistency, but not rates.

Knowing  $\beta_a$  would let us determine the effective sample size of time series  $Y_{1:n}$ . In effect, having  $n$  dependent-but-mixing data points is like having  $\mu < n$  independent ones. Once we determine the correct  $\mu$ , we can (as we will now show) use concentration results for IID data like those in [Theorem 2.1](#) and [Theorem 2.5](#) with small corrections.

## 4 Risk bounds

With the relevant background in place, we can put the pieces together to derive our results. We use  $\beta$ -mixing to find out how much information is in the data and VC dimension to measure the capacity of the state-space model's prediction functions. The result is a bound on the generalization error of the chosen function  $\hat{f}$ . After slightly modifying the definition of “risk” to fit the time-series forecasting scenario, and stating necessary technical assumptions, we derive risk bounds for wide classes of economic forecasting models.

### 4.1 Setup and assumptions

We observe a finite subsequence of random vectors  $Y_{1:n}$  from a process  $Y_\infty$  defined on a probability space  $(\Omega, \Sigma, \mathbb{P}_\infty)$ , with  $Y_i \in \mathbb{R}^p$ . We make the following assumption on the process.

**Assumption A.**  $\mathbb{P}_\infty$  is a stationary,  $\beta$ -mixing process with mixing coefficients<sup>5</sup>  $\beta_a$ ,  $\forall a > 0$ .

Under stationarity, the marginal distribution of  $Y_t$  is the same for all  $t$ . We deal mainly with the joint distribution of  $Y_{1:n+1}$ , where we observe the first  $n$  observations and try predicting  $Y_{n+1}$ . For the rest of this paper, we will call this joint distribution  $\mathbb{P}$ . Our results extend to predicting more than one step ahead, but the notation becomes cumbersome.

We must define generalization error and training error slightly differently for time series than in the IID setting. Using the same notion of loss functions as before, we consider prediction functions  $f : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^p$

**Definition 4.1** (Time series risk).

$$R_n(f) := \mathbb{E} \left[ \ell(Y_{n+1} - f(Y_{1:n})) \right]. \quad (25)$$

The expectation is taken with respect to the joint distribution  $\mathbb{P}$  and therefore depends on  $n$ . The function  $f$  may use some or all of the past to generate predictions. A function using only the most recent  $d$  observations as inputs will be said to have *fixed memory* of length  $d$ . Other functions have *growing memory*, i.e.,  $f$  may use all the previous data to predict the next data point. This incongruity makes the notation for time series training error somewhat problematic.

We will define the training error with a subscript  $i \in \mathbb{N}$  on  $f$  within the summation. Strictly speaking, there is only one function  $f$  which we are using to make forecasts. In typical fixed memory settings — standard VAR forecasting models and so on —  $f_i = f_j = f$  for all  $i, j \in \mathbb{N}$ . But for models with growing memory, a fixed forecasting method — an ARMA model, DSGE,<sup>6</sup> or linear state-space model — will use all of the past to make predictions, so the dimension of the domain changes with  $i$ . We write the risk of  $f$  as a single function, because, once we parameterize a forecasting method, an entire sequence of forecasting functions  $f_1, f_2, \dots$  is determined.

**Definition 4.2** (Time series training error).

$$\hat{R}_n(f) := \frac{1}{n-d-1} \sum_{i=d}^{n-1} \ell(Y_{i+1} - f_i(Y_{1:i})). \quad (26)$$

<sup>5</sup>In order to apply the results, one must either know  $\beta_a$  for some  $a$  or be able to estimate it with sufficient precision and accuracy. McDonald et al. [34] shows how to estimate the mixing coefficients non-parametrically, based on a single sample from the process.

<sup>6</sup>A DSGE is a nonlinear system of expectational difference equations, so estimating the parameters is nontrivial. Likelihood methods typically work by finding a linear approximation using Taylor expansions and the Kalman filter, though increasingly complex nonlinear methods are now intensely studied. See for instance DeJong and Dave [13], Fernández-Villaverde [20] or DeJong et al. [15]



In order to make use of this single definition of training error, we let  $d \geq 0$ . In fixed memory cases — say an AR(2) —  $d$  has an obvious meaning, while with growing memory,  $d = 0$  is allowed.

To control the generalization error for time series forecasting, we make one final assumption, about the possible magnitude of the losses. Specifically, we weaken the bounded loss assumption we used in §2 to allow for unbounded loss as long as we retain some control on moments of the loss.

**Assumption B.** Assume that for all  $f \in \mathcal{F}$

$$Q_n(f) := \sqrt{\mathbb{E}_{\mathbb{P}} \left[ \ell(Y_{n+1} - f(Y_{1:n}))^2 \right]} \leq M < \infty. \quad (27)$$

Assumption B is still quite general, allowing even some heavy tailed distributions.

## 4.2 Fixed memory

We can now state our results giving finite sample risk bounds for the problem of time series forecasting. We begin with the fixed memory setting; the next section will allow the memory length to grow.

**Theorem 4.3.** Suppose that Assumption A and Assumption B hold, that the model class  $\mathcal{F}$  has a fixed memory length  $d < n$ , and that we have a sample  $\mathbf{Y}_1^n$ . Let  $\mu$  and  $a$  be integers such that  $2\mu + d \leq n$ . Then, for all  $\epsilon > 0$ ,

$$\begin{aligned} & \mathbb{P} \left( \sup_{f \in \mathcal{F}} \frac{R_n(f) - \hat{R}_n(f)}{Q_n(f)} > \epsilon \right) \\ & \leq 8(2\mu + 1)^{\text{vcd}(\mathcal{F})} \exp \left\{ -\frac{\mu \exp \left( W \left( -\frac{2\epsilon^2}{e^4} \right) + 4 \right)}{4} \right\} + 2\mu\beta_{a-d}, \end{aligned} \quad (28)$$

where  $W(\cdot)$  is the Lambert W function.

The implications of this theorem are considerable. Given a finite sample of length  $n$ , we can say that with high probability, future prediction errors will not be much larger than our observed training errors. It makes no difference whether the model is correctly specified. This stands in stark contrast to model selection tools like AIC or BIC which appeal to asymptotics. Moreover, given a model class  $\mathcal{F}$ , we can say exactly how much data we need to have good control of the prediction risk. As the effective data size increases, the training error is a better and better estimate of the generalization error, uniformly over all of  $\mathcal{F}$ .

The Lambert W function in the exponential term deserves some explanation. The Lambert W function is defined as the inverse of  $f(w) = w \exp w$  (cf. Corless et al. [9]). A strictly, but only slightly, worse bound can be achieved by noting that

$$\exp \left( W \left( -\frac{2\epsilon^2}{e^4} \right) + 4 \right) \leq \frac{\epsilon^{8/3}}{4^{2/3}} \quad (29)$$

for all  $\epsilon \in [0, 1]$ .

The difference between expected and empirical risk is only interesting when  $R_n(f)$  exceeds  $\hat{R}_n(f)$ . Due to the supremum, events where the training error exceeds the expected risk are irrelevant. Therefore, we are only concerned with  $0 \leq \hat{R}_n(f) \leq R_n(f)$ . Of course, as discussed in Section 2, for most estimation procedures,  $f$  is chosen to make  $\hat{R}_n(f)$  as small as possible.

One way to understand this theorem is to visualize the tradeoff between confidence  $\epsilon$  and effective data  $\mu$ . Consider, by way of illustration, what happens when  $\text{vcd}(\mathcal{F}) = 1$ ,  $\beta_a = 0$ , and  $M = 1$ . Then (28) and (29) become

$$\mathbb{P} \left( \sup_{f \in \mathcal{F}} R_n(f) - \hat{R}_n(f) > \epsilon \right) \leq 8 \exp \left\{ \log(2\mu + 1) - \frac{\mu\epsilon^{8/3}}{4^{5/3}} \right\} \quad (30)$$

Our goal is to minimize  $\epsilon$ , thereby ensuring that the relative difference between the expected risk and the training risk is small. At the same time we want to minimize the right side of the bound so that the probability of “bad” outcomes — samples where the difference in risks exceeds  $\epsilon$  — is small. Of course

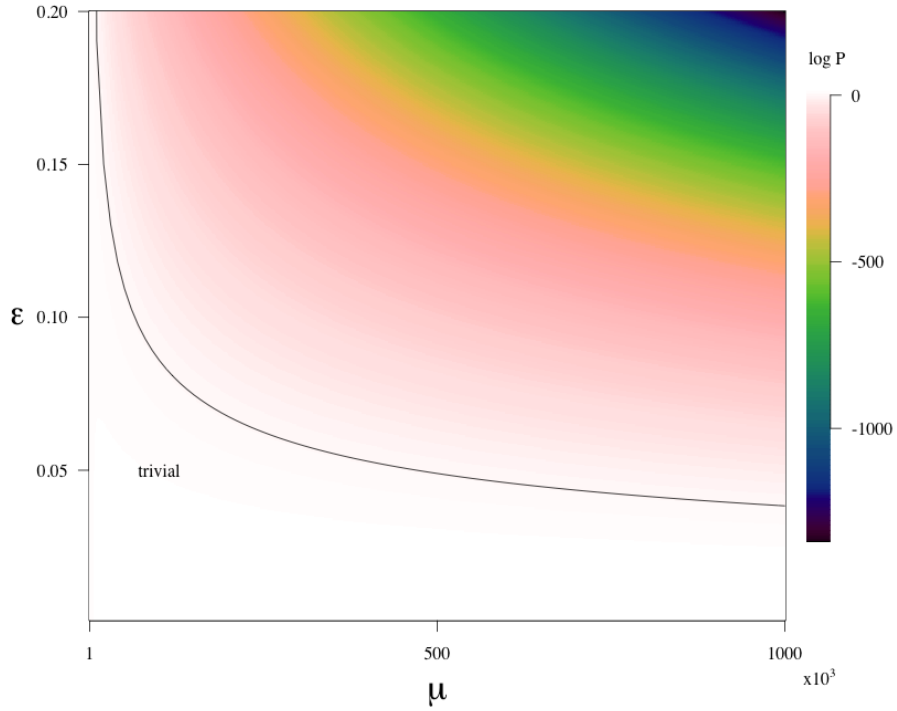


Figure 1: Visualizing the tradeoff between confidence ( $\epsilon$ ,  $y$ -axis) and effective data ( $\mu$ ,  $x$ -axis). The black curve indicates the region where the bound becomes trivial. Below this line, the probability is bounded by 1. Darker colors indicate lower probability of the “bad” event — that the difference in risks exceeds  $\epsilon$ . The colors correspond to the natural logarithm of the bound on this probability.

we want to do this with as little data as possible, but the smaller we take  $\epsilon$ , the larger we must take  $\mu$  to compensate. We depict this tradeoff in Figure 1.

The figure is structured so that movement toward the origin is preferable. We have tighter control on the difference in risks with less data. But moving in that direction leads to an increased probability of the bad event — that the difference in risks exceeds  $\epsilon$ . The bound becomes trivial below the solid black line (the bad event occurs with probability no larger than one). The desire for the bad event to occur with low probability forces the decision boundary to the upper right.

Another way to interpret the plot is as a set of indifference curves. Anywhere in the same color region is equally desirable in the sense that the probability of equally bad events is the same. So if we had a budget constraint trading  $\epsilon$  and data (i.e. a line with negative slope), we could optimize within the budget set to find the lowest probability allowable.

Before we prove Theorem 4.3, we will state a corollary which puts the same result in a form that is sometimes easier to use.

**Corollary 4.4.** *Under the conditions of Theorem 4.3, for any  $f \in \mathcal{F}$ , the following bound holds with probability at least  $1 - \eta$ , for all  $\eta > 2\mu\beta_{a-d}$ :*

$$R_n(f) \leq \hat{R}_n(f) + M \sqrt{\frac{\mathcal{E}(4 - \log \mathcal{E})}{2}}, \quad (31)$$

with

$$\mathcal{E} = \frac{4 \text{VCD}(\mathcal{F}) \log(2\mu + 1) + \log 8/\eta'}{\mu}, \quad (32)$$

and  $\eta' = \eta - 2\mu\beta_{a-d}$ .

We now prove both [Theorem 4.3](#) and [Corollary 4.4](#) to provide the reader with some intuition for the types of arguments necessary. We defer proof of the remainder of the theorems in this section to the appendix.

*Proof of Theorem 4.3 and Corollary 4.4.* The first step is to move from the actual sample size  $n$  to the effective sample size  $\mu$  which depends on the  $\beta$ -mixing behavior. Let  $a$  and  $\mu$  be non-negative integers such that  $2a\mu + d \leq n$ . Now divide  $\mathbf{Y}_1^n$  into  $2\mu$  blocks, each of length  $a$ , ignoring the remainder. Identify the blocks as follows:

$$U_j = \{Y_i : 2(j-1)a + 1 \leq i \leq (2j-1)a\}, \quad (33)$$

$$V_j = \{Y_i : (2j-1)a + 1 \leq i \leq 2ja\}. \quad (34)$$

Let  $\mathbf{U}$  be the sequence of odd blocks  $U_j$ , and let  $\mathbf{V}$  be the sequence of even blocks  $V_j$ . Finally, let  $\mathbf{U}'$  be a sequence of blocks which are mutually independent and such that each block has the same distribution as a block from the original sequence. That is construct  $U'_j$  such that

$$\mathcal{L}(U'_j) = \mathcal{L}(U_j) = \mathcal{L}(U_1), \quad (35)$$

where  $\mathcal{L}(\cdot)$  means the probability law of the argument.

Let  $\hat{R}_{\mathbf{U}}(f)$ ,  $\hat{R}_{\mathbf{U}'}(f)$ , and  $\hat{R}_{\mathbf{V}}(f)$  be the empirical risk of  $f$  based on the block sequences  $\mathbf{U}$ ,  $\mathbf{U}'$ , and  $\mathbf{V}$  respectively. Clearly  $\hat{R}_n(f) = \frac{1}{2}(\hat{R}_{\mathbf{U}}(f) + \hat{R}_{\mathbf{V}}(f))$ . Then,

$$\mathbb{P} \left( \sup_{f \in \mathcal{F}} \frac{R_n(f) - \hat{R}_n(f)}{Q_n(f)} > \epsilon \right) \quad (36)$$

$$\begin{aligned} &= \mathbb{P} \left( \sup_{f \in \mathcal{F}} \left[ \frac{R_n(f) - \hat{R}_{\mathbf{U}}(f)}{2Q_n(f)} + \frac{R_n(f) - \hat{R}_{\mathbf{V}}(f)}{2Q_n(f)} \right] > \epsilon \right) \\ &\leq \mathbb{P} \left( \sup_{f \in \mathcal{F}} \frac{R_n(f) - \hat{R}_{\mathbf{U}}(f)}{Q_n(f)} + \sup_{f \in \mathcal{F}} \frac{R_n(f) - \hat{R}_{\mathbf{V}}(f)}{Q_n(f)} > 2\epsilon \right) \end{aligned} \quad (37)$$

$$\leq \mathbb{P} \left( \sup_{f \in \mathcal{F}} \frac{R_n(f) - \hat{R}_{\mathbf{U}}(f)}{Q_n(f)} > \epsilon \right) + \mathbb{P} \left( \sup_{f \in \mathcal{F}} \frac{R_n(f) - \hat{R}_{\mathbf{V}}(f)}{Q_n(f)} > \epsilon \right) \quad (38)$$

$$= 2\mathbb{P} \left( \sup_{f \in \mathcal{F}} \frac{R_n(f) - \hat{R}_{\mathbf{U}}(f)}{Q_n(f)} > \epsilon \right). \quad (39)$$

Now, apply Lemma 4.1 in Yu [56] (reproduced as [Lemma A.1](#) in [Section A](#)) to the of the event  $\left\{ \sup_{f \in \mathcal{F}} \frac{R_n(f) - \hat{R}_{\mathbf{U}}(f)}{Q_n(f)} > \epsilon \right\}$ . This allows us to move from statements about dependent blocks to statements about independent blocks with a slight correction. Therefore we have,

$$\begin{aligned} &2\mathbb{P} \left( \sup_{f \in \mathcal{F}} \frac{R_n(f) - \hat{R}_{\mathbf{U}}(f)}{Q_n(f)} > \epsilon \right) \\ &\leq 2\mathbb{P} \left( \sup_{f \in \mathcal{F}} \frac{R_n(f) - \hat{R}_{\mathbf{U}'}(f)}{Q_n(f)} > \epsilon \right) + 2(\mu - 1)\beta_{a-d}, \end{aligned} \quad (40)$$

where the probability on the right is for the  $\sigma$ -field generated by the independent block sequence  $\mathbf{U}'$ . Therefore,

$$\begin{aligned} &2\mathbb{P} \left( \sup_{f \in \mathcal{F}} \frac{R_n(f) - \hat{R}_{\mathbf{U}'}(f)}{Q_n(f)} > \epsilon \right) \\ &\leq 8(2\mu + 1)^{\text{vcd}(\mathcal{F})} \exp \left\{ -\frac{\mu \exp \left( W \left( -\frac{2\epsilon^2}{e^4} \right) + 4 \right)}{4} \right\} \end{aligned} \quad (41)$$

where we have applied Theorem 7 in Cortes et al. [10] (reproduced as Lemma A.2) to bound the independent blocks  $\mathbf{U}'$ .

To prove the corollary, set the right hand side of (41) to  $\eta$ , take  $\eta' = \eta - 2(\mu - 1)\beta_{a-d}$ , and solve for  $\epsilon$ . We get that for all  $f \in \mathcal{F}$ , with probability at least  $1 - \eta$ ,

$$\frac{R_n(f) - \hat{R}_n(f)}{Q_n(f)} \leq \epsilon. \quad (42)$$

Solving the equation

$$\eta' = 8(2\mu + 1)^h \exp \left\{ -\frac{\mu \exp \left( W \left( -\frac{2\epsilon^2}{e^4} \right) + 4 \right)}{4} \right\} \quad (43)$$

implies

$$\epsilon = M \sqrt{\frac{\mathcal{E}(4 - \log \mathcal{E})}{2}} \quad (44)$$

with

$$\mathcal{E} = \frac{4 \text{VCD}(\mathcal{F}) \log(2\mu + 1) + \log 8/\eta'}{\mu}. \quad (45)$$

■

The only obstacle to the use of Theorem 4.3 is knowledge of  $\text{VCD}(\mathcal{F})$ . For some models, the VC dimension can be calculated explicitly.

**Theorem 4.5.** *For the class of  $AR(d)$  models,  $\mathcal{F}_{AR}(d)$ ,*

$$\text{VCD}(\mathcal{F}_{AR}(d)) = d + 1. \quad (46)$$

*For the class of  $VAR(d)$  models with  $k$  time series,  $\mathcal{F}_{VAR}(k, d)$ ,*

$$\text{VCD}(\mathcal{F}_{VAR}(k, d)) = kd + 1. \quad (47)$$

Theorem 4.5 applies equally to Bayesian VARs. However, this is likely too conservative as the prior tends to restrict the effective complexity of the function class.<sup>7</sup>

### 4.3 Growing memory

Most macroeconomic forecasting model classes have growing rather than fixed-length memories. These model classes include dynamic factor models, ARMA models, and linearized dynamic stochastic general equilibrium models. However, all of these models have the property that forecasts are linear functions of past observations, and, moreover, the weight placed on the past generally shrinks exponentially. These properties let us get bounds similar to our previous results.

Any linear predictors with growing memory can be put in the following form ( $1 \leq d < n$ ):

$$\hat{Y}_{d+1:n+1} = \mathbf{B}Y_{1:n} \quad (48)$$

---

<sup>7</sup>Here we should mention that these risk bounds are frequentist in nature. We mean that if one treats Bayesian methods as a regularization technique and predicts with the posterior mean or mode, then our results hold. However, from a subjective Bayesian perspective, our results add nothing since all inference can be derived from the posterior. For further discussion of the frequentist risk properties of Bayesian methods under mis-specification, see for example Kleijn and van der Vaart [29], Müller [41] or Shalizi [47]

where

$$\mathbf{B} = \begin{bmatrix} b_{d,1} & \cdots & b_{d,d} & & & 0 \\ b_{d+1,1} & \cdots & b_{d+1,d} & b_{d+1,d+1} & & \\ \vdots & & \vdots & & \ddots & \\ b_{n,1} & \cdots & b_{n,d} & b_{n,d+1} & \cdots & b_{n,n} \end{bmatrix}. \quad (49)$$

With this notation, we can prove the following about growing memory linear predictors.

**Theorem 4.6.** *Suppose that [Assumption A](#) and [Assumption B](#) hold, and that the model class  $\mathcal{F}$  is linear in the data, with growing memory. Further assume that the loss function  $\ell$  satisfies the following conditions:*

1. *for some  $\Delta > 0$ ,  $\ell(y + y') \leq \Delta(\ell(y) + \ell(y'))$  (modified triangle inequality).*
2.  *$\ell(yy') \leq \ell(y)\ell(y')$  (sub-multiplication).*

*Given a time-series of length  $n$ , fix some  $1 \leq d < n$ , and let  $\mu$  and  $a$  be integers such that  $2\mu a + d \leq n$ . Then the following bound holds simultaneously for all  $f \in \mathcal{F}$ :*

$$\begin{aligned} & \mathbb{P} \left( \sup_{f \in \mathcal{F}} \frac{R_n(f) - \hat{R}_n(f) - \delta_d(f)}{Q_n(f)} > \epsilon \right) \\ & \leq 8(2\mu + 1)^h \exp \left\{ -\frac{\mu \exp \left( W \left( -\frac{2\epsilon^2}{e^4} \right) + 4 \right)}{4} \right\} + 2\mu\beta_{a-d}, \end{aligned} \quad (50)$$

where

$$\delta_d(f) = \Delta^2 \mathbb{E}[\ell(Y_1)] \sum_{j=1}^{n-d-1} \ell(b_{n,j}) + \frac{\Delta}{n-d-1} \sum_{i=d+1}^{n-1} \ell \left( \sum_{j=1}^{i-d} b_{i,j} y_j \right). \quad (51)$$

We should clarify the conditions on the loss function, and the role of the approximation term.

The assumptions on the loss function are quite mild. Both conditions are satisfied for any norm: the triangle inequality holds with  $\Delta = 1$  (by the definition of “norm”), and sub-multiplication holds by the Cauchy-Schwarz inequality. Thus the assumptions hold when, for instance, vector-valued predictions have their accuracy measured using matrix norms. Likewise, absolute error loss ( $\ell(y - y') = |y - y'|$ ) satisfies both conditions with  $\Delta = 1$ , while squared error loss satisfies the conditions with  $\Delta = 2$ .

The  $\delta_d(f)$  term arises from taking a fixed-memory approximation, of length  $d$ , to predictors with growing memory. As will become clear in the proof, we make this approximation to apply the previous theorem, but it involves a trade-off. As  $d \nearrow n$ ,  $\delta_d(f) \searrow 0$ , but this drives  $\mu \searrow 0$ , resulting in fewer effective training points whereas smaller  $d$  has the opposite effect. Also,  $\delta_d(f)$  depends on  $\mathbb{E}[\ell(Y_1)]$  which is not necessarily desirable. However, [Assumption B](#) has the consequence that  $\mathbb{E}[\ell(Y_1)] \leq M < \infty$ .

**Corollary 4.7.** *Given a sample  $\mathbf{Y}_1^n$  such that [Assumption A](#) and [Assumption B](#) hold, suppose that the model class  $\mathcal{F}$  is linear in the data and has growing memory. Fix some  $1 \leq d < n$ . Then, for any  $f \in \mathcal{F}$ , the following bound holds with probability at least  $1 - \eta$ ,*

$$R_n(f) \leq \hat{R}_n(f) + \delta_d(f) + M \sqrt{\frac{\mathcal{E}(4 - \log \mathcal{E})}{2}}, \quad (52)$$

where  $\mathcal{E}$  and  $\eta$  are as in [Theorem 4.3](#).

To apply [Theorem 4.6](#), we specialize to linear Gaussian state-space models, where we can calculate  $\delta_d(f)$  directly, and demonstrate that it will behave well as  $n$  grows. Such models are not, unfortunately, universal, but all of the most common macroeconomic forecasting models — including dynamic factor models, ARMA models, GARCH models, and even linearized DSGEs — have linear-Gaussian state-space representations.

The general specification of a linear Gaussian state-space model,  $\mathcal{F}_{SS}$ , is

$$\begin{aligned} y_t &= Z\alpha_t + \epsilon_t, & \epsilon_t &\sim \mathcal{N}(0, H), \\ \alpha_{t+1} &= T\alpha_t + \eta_{t+1}, & \eta_t &\sim \mathcal{N}(0, Q), \\ & & \alpha_1 &\sim \mathcal{N}(a_1, P_1). \end{aligned} \quad (53)$$

We make no assumptions about the sizes of the parameter matrices  $Z$ ,  $T$ ,  $H$ ,  $Q$ ,  $a_1$ , or  $P_1$ , but we do require stationarity. This amounts to forcing the eigenvalues of  $T$  to lie inside the complex unit circle. Stationarity ensures that  $\delta_d(f)$  will be bounded as well as conforming to our assumptions about the data generating process.

To forecast using  $\mathcal{F}_{SS}$ , one uses the Kalman filter (Durbin and Koopman [18], Kalman [27]). To estimate the unknown parameter matrices, we either: (1) maximize the likelihood returned by the filter; or (2) use the EM algorithm, alternating between running the Kalman filter (the E-step) and maximizing the conditional likelihood by least squares (the M-step). (Bayesian estimation works like EM, replacing the M-step with Bayesian updating.) Either way, one can show [18] that given the parameter matrices, the (maximum *a posteriori*) forecast of  $y_t$  is given by

$$\hat{y}_{t+1} = Z \sum_{j=1}^{t-1} \prod_{i=j+1}^t L_i K_j y_j + Z K_t y_t + Z \prod_{i=1}^t L_i a_1 \quad (54)$$

where

$$\begin{aligned} F_t &= (Z P_t Z' + H)^{-1}, & K_t &= T P_t Z' F_t, \\ L_t &= T - K_t Z, & P_{t+1} &= T P_t L_t' + Q. \end{aligned} \quad (55)$$

This yields the form of  $\delta_d(f)$  for linear state-space models. We therefore have the following corollary to [Theorem 4.6](#).

**Corollary 4.8.** *Let  $\mathcal{F}$  correspond to a state-space model as in (53), and fix  $1 < d < n$ . Then the following bound holds simultaneously for all  $f \in \mathcal{F}$ : with probability at least  $1 - \eta$ ,*

$$R_n(f) \leq \hat{R}_n(f) + \delta_d(f) + M \sqrt{\frac{\mathcal{E}(4 - \log \mathcal{E})}{2}}, \quad (56)$$

where  $\mathcal{E}$  is as in [Theorem 4.3](#), and

$$\begin{aligned} \delta_d(f) &= \Delta^2 \mathbb{E}[\ell(Y_1)] \sum_{j=1}^{n-d} \ell \left( \prod_{i=j+1}^n L_i K_j \right) \\ &\quad + \frac{\Delta}{n-d-1} \sum_{t=d+1}^{n-1} \ell \left( \sum_{j=1}^{t-d} \prod_{i=j+1}^t L_i K_j y_j \right). \end{aligned} \quad (57)$$

It is simple to compute  $\delta_d(f)$  using Kalman filter output, so the corollary lets us compute risk bounds for common macroeconomic forecasting models.

## 5 Bounds in practice

We now show how the theorems of the previous section can be used both to quantify prediction risk and to select models. We first estimate a simple stochastic volatility model using IBM return data and calculate the bound for the predicted volatility using [Corollary 4.8](#). Then we show how the same methods can be used for typical macroeconomic forecasting models.



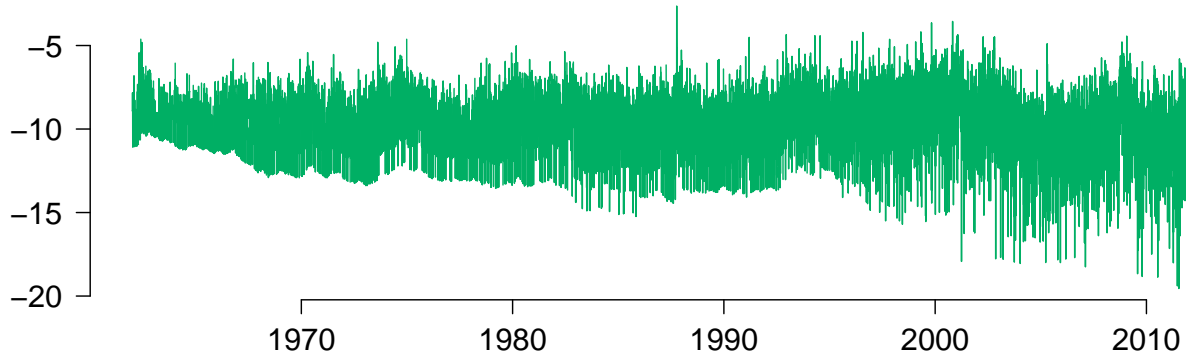


Figure 2: This figure plots daily volatility (squared log returns) for IBM from 1962–2011.

### 5.1 Stochastic volatility model

We estimate a standard stochastic volatility model using daily log returns for IBM from January 1962 until October 2011 —  $n = 12541$  observations. Figure 2 shows the squared log-return series.

The model we investigate is

$$y_t = \sigma z_t \exp(\rho_t/2), \quad z_t \sim N(0, 1), \quad (58)$$

$$\rho_{t+1} = \phi \rho_t + w_t, \quad w_t \sim N(0, \sigma_w^2), \quad (59)$$

where the disturbances  $z_t$  and  $w_t$  are mutually and serially independent. Following Harvey et al. [24], we linearize this non-linear model as follows:

$$\log y_t^2 = \kappa + \frac{1}{2} \rho_t + \xi_t, \quad (60)$$

$$\xi_t = \log z_t^2 - \mathbb{E}[\log z_t^2], \quad (61)$$

$$\kappa = \log \sigma^2 + \mathbb{E}[\log z_t^2]. \quad (62)$$

The noise term  $\xi_t$  is no longer Gaussian, but the Kalman filter will still give the minimum-mean-squared-error linear estimate of the variance sequence  $\rho_{1:n+1}$ . The observation variance is now  $\pi^2/2$ .

To match the data to the model, let  $y_t$  be the log returns and remove 688 observations where the return was 0 (i.e., the price did not change from one day to the next). Using the Kalman filter, the negative log likelihood is given by

$$\mathcal{L}(Y_{1:n}|\kappa, \phi, \sigma_\rho^2) \propto \sum_{t=1}^n \log F_t + v_t^2 F_t^{-1}. \quad (63)$$

Minimizing this gives estimates  $\kappa = -9.62$ ,  $\phi = 0.996$ , and  $\sigma_w^2 = 0.003$ . Taking the  $\ell(y, y') = (y - y')^2$  gives training error  $\hat{R}_n(f) = 3.333$ .

To actually calculate the bound, we need a few more values. First, using the methods in McDonald et al. [34, 35], we can estimate  $\beta_8 = 0.017$ . For  $a > 8$ , the optimal point estimate of  $\beta_a$  is 0. While this is presumably an underestimate, we will take  $\beta_a = 0$  for  $a > 8$ . For the upper bound in Assumption B, we use  $M = \sqrt{2}$ .

Combining these values with the VC dimension for the stochastic volatility model, we can bound the prediction risk. For  $d = 2$ , the VC dimension can be no larger than 3. Finally, taking  $\mu = 538$ ,  $a = 11$ ,  $d = 2$ , and  $\mathbb{E}[Y_1^2] = 1$ , we get that  $\delta_2(f) = 0.60 + 2.13 = 2.73$ . The result is the bound

$$R_n(f) \leq 7.04 \quad (64)$$

with probability at least 0.85. In other words, the bound is much larger than the training error, but this is to be expected: the data are highly dependent, so the large  $n$  translates into a relatively small effective sample size  $\mu$ .

Table 1: This table shows the training error and risk bounds for 3 models. AIC is given as the difference from predicting with the global mean (the smaller the value, the more support for that model).

Model	Training error	AIC-Baseline	Risk bound ( $1 - \eta > 0.85$ )
SV	3.33	-2816	7.04
AR(2)	3.54	-348	4.52
Mean	3.65	0	4.29

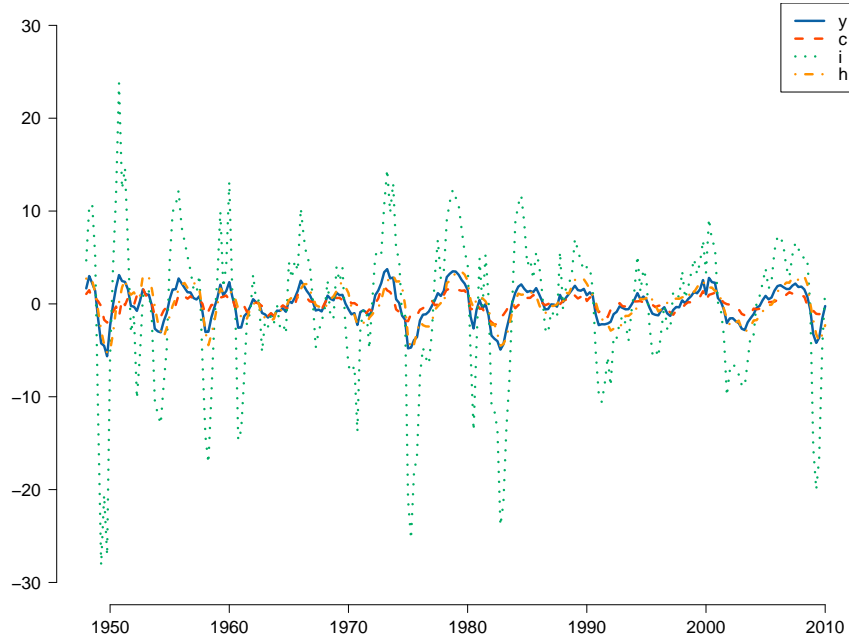


Figure 3: Time series used to estimate the RBC model. These are quarterly data from 1948:I until 2010:I. The blue line is GDP (output), the red line is consumption, the green line is investment, and the orange line is hours worked. These data are plotted as percentage deviations from trend as discussed in [Section C](#).

For comparison, we also computed the bound for forecasts produced with an AR(2) model (with intercept) and with the global mean alone. In the case of the mean, we take  $\mu = 658$  and  $a = 9$  since in this case,  $d = 0$ . The results are shown in [Table 1](#). The stochastic volatility model reduces the training error by 5% relative to predicting with the mean, an improvement which is marginal at best. But the resulting risk bound clearly demonstrates that given the small effective sample size, even this gain may be spurious: it is likely that the stochastic volatility model is simply over-fitting.

## 5.2 Real business cycle model

In this section, we will discuss the methodology for applying risk bounds to the forecasts generated by the real business cycle (RBC) model. This is a standard tool in macroeconomic forecasting. For a discussion of the RBC model and the standard methods used to bring the model to the data, see, for example DeJong and Dave [13], DeJong et al. [14], Fernández-Villaverde [20], Kydland and Prescott [30], Romer [46], Sims [49], Smets and Wouters [50].

To estimate the parameters of this model, we use four data series. These are GDP  $y_t$ , consumption  $c_t$ , investment  $i_t$ , and hours worked  $n_t$ . (The data from the Federal Reserve Economic Database, FRED.) The series we use are shown in [Figure 3](#).

The basic idea of the estimation is to transform the model from an inter-temporal optimization form into a state space model. This leads to a linear, Gaussian state-space model with four observed variables (listed above), and two unobserved state variables. The mapping from parameters of the optimization problem to

parameters of the state-space model is nonlinear, but, for each parameter setting, the Kalman filter returns the likelihood, so that likelihood methods are possible. As the data are uninformative about many of the parameters, we estimate by maximizing a penalized likelihood, rather than a simple likelihood. Then the Kalman filter produces in-sample forecasts which are linear in past values of the data, so that we could potentially apply the growing memory bound.

For macroeconomic time series, there is not enough data to give nontrivial bounds, regardless of the mixing coefficients or the size of the finite memory approximation. Figure 3 shows  $n = 249$  observations. The minimal possible finite approximation model is a VAR with one lag and four time series, which, by Theorem 4.5, has VC dimension 5. In this case, since we are dealing with vector valued forecasts, we take  $\ell(y - y') = \|y - y'\|_2$ . We assume that the Assumption B is satisfied with  $M = 0.1$  and demand confidence 0.85 ( $\eta = 0.15$ ),

Again, using the methods of McDonald et al. [34, 35], we can estimate the  $\beta$ -mixing coefficients of the macroeconomic data set. The result is a point estimate  $\beta_4 = 0$ . Assuming that this is approximately accurate (0 is of course an underestimate), this suggests that the effective size of the macroeconomic data set is no more than about  $\mu = 31$ , much smaller than  $n = 249$ . To calculate the bound, we assume that  $\mathbb{E}[\|Y\|_2] < 0.1$ . Since the loss function is a norm, then  $\Delta = 1$ . The training error of the fitted RBC model is  $\hat{R}_n(f) = 0.00059$ . Thus our bound is given by

$$R_n(f) \leq \hat{R}_n(f) + \delta_1(f) + \text{penalty} = 0.00059 + 0.18 + 3.07 = 3.26. \quad (65)$$

The bound here is four orders of magnitude larger than the training error. If the bound is tight, then this suggests that the training error severely underestimates the true prediction risk. Of course, this should not be too surprising since the RBC model has 11 parameters and we are trying to get confidence intervals using only 31 effective data points.

In some sense, the empirical results in this section seem slightly unreasonable. Since the results are only upper bounds, it is important to get an idea as to how tight they may be. We address this issue in the next section.

## 6 Properties of our results

In the previous section, we showed that the upper bound for the risk of standard macroeconomic forecasting models may be large. This of course raises the question “How tight are these bounds?” We address this question next and then discuss how to use the bounds for model selection.

### 6.1 How tight are the bounds?

Here we give some idea of how tight the bounds presented in Section 4 are. Call  $\hat{f}_{erm}$  is the function that minimizes the training error (or penalized training error) over  $\mathcal{F}$ , and

$$f^* = \operatorname{argmin}_{f \in \mathcal{F}} R_n(f) \quad (66)$$

is the minimizer of the true risk (“pseudo-truth”), i.e. the best-predicting function in  $\mathcal{F}$ . We call

$$L_n(\Pi) := \inf_{\mathbb{P} \in \Pi} \mathbb{E}_{\mathbb{P}}[R_n(\hat{f}_{erm}) - R_n(f^*)] = \inf_{\mathbb{P} \in \Pi} \mathbb{E}_{\mathbb{P}}[R_n(\hat{f}_{erm})] - R_n(f^*) \quad (67)$$

the “oracle loss”; it describes how well empirical risk minimization works relative to the best possible predictor  $f^*$  over the worst distribution  $\mathbb{P}$ . Vapnik [52] shows that for classification and IID data, for sufficiently large  $n$ , there exist constants  $c$  and  $C$  such that

$$c\sqrt{\frac{\text{VCD}(\mathcal{F})}{n}} \leq L_n(\Pi) \leq C\sqrt{\frac{\text{VCD}(\mathcal{F}) \log n}{n}}, \quad (68)$$

where  $\Pi$  is the class of all distributions satisfying  $\mathbb{P}(\ell(y - y') > K) = 0$ . In other words, for IID data, the best we can hope to do is a rate of  $O\left(\sqrt{\frac{\text{VCD}(\mathcal{F})}{n}}\right)$  and prediction methods which perform worse than

$O\left(\sqrt{\frac{\text{VCD}(\mathcal{F}) \log n}{n}}\right)$  are inefficient. We will derive similar bounds for the  $\beta$ -mixing setting. First, we need a slightly different version of [Theorem 4.3](#).

**Theorem 6.1.** *Suppose that  $\ell(y - y') < K$ , that [Assumption A](#) holds, and that  $\mathcal{F}$  has a fixed memory length  $d < n$ . Let  $\mu$  and  $a$  be integers such that  $2\mu + d \leq n$ . Then, for all  $\epsilon > 0$ ,*

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |R_n(f) - \hat{R}_n(f)| > \epsilon\right) \leq 8(2\mu + 1)^{\text{VCD}(\mathcal{F})} \exp\left\{-\frac{\mu\epsilon^2}{K_1^2}\right\} + 2\mu\beta_{a-d}. \quad (69)$$

where  $K_1$  depends only on  $K$ .

The proof of [Theorem 6.1](#) is exactly like that for [Theorem 4.3](#).

**Assumption C.** *The time series  $Y_\infty$  is exponentially (or geometrically)  $\beta$ -mixing, i.e.*

$$\beta_a = c_1 \exp(-c_2 a^\kappa) \quad (70)$$

for some constants  $c_1, c_2, \kappa$ .

**Theorem 6.2.** *Suppose  $\ell(y - y') < K$  and that [Assumption C](#) holds. Then, for sufficiently large  $n$ , there exist constants  $c$  and  $C$ , independent of  $n$  and  $\text{VCD}(\mathcal{F})$ , such that*

$$c\sqrt{\frac{\text{VCD}(\mathcal{F})}{n}} \leq L_n(\Pi) \leq C\sqrt{\frac{\text{VCD}(\mathcal{F}) \log n}{n^{\kappa/(1+\kappa)}}}. \quad (71)$$

*Proof.* [Theorem 6.1](#) implies that simultaneously

$$\begin{aligned} & \mathbb{P}\left(|R_n(\hat{f}_{erm}) - \hat{R}_n(\hat{f}_{erm})| > \epsilon\right) \\ & \leq 8(2\mu + 1)^{\text{VCD}(\mathcal{F})} \exp\left\{-\frac{\mu\epsilon^2}{K_1^2}\right\} + 2(\mu - 1)\beta_{a-d} \end{aligned} \quad (72)$$

and

$$\begin{aligned} & \mathbb{P}\left(|R_n(f^*) - \hat{R}_n(f^*)| > \epsilon\right) \\ & \leq 8(2\mu + 1)^{\text{VCD}(\mathcal{F})} \exp\left\{-\frac{\mu\epsilon^2}{K_1^2}\right\} + 2(\mu - 1)\beta_{a-d}. \end{aligned} \quad (73)$$

Since  $\hat{R}_n(\hat{f}_{erm}) - \hat{R}_n(f^*) \leq 0$ , then

$$\begin{aligned} & \mathbb{P}\left(|R_n(\hat{f}_{erm}) - R_n(f^*)| > 2\epsilon\right) \\ & \leq 8(2\mu + 1)^{\text{VCD}(\mathcal{F})} \exp\left\{-\frac{\mu\epsilon^2}{K_1^2}\right\} + 2(\mu - 1)\beta_{a-d}. \end{aligned} \quad (74)$$

Letting  $Z = |R_n(\hat{f}_{erm}) - R_n(f^*)|$ ,  $k_1 = 8(2\mu + 1)^{\text{VCD}(\mathcal{F})}$ , and  $k_2 = 1/K_1^2$  and ignoring constants,

$$\mathbb{E}[Z^2] \leq s + k'_1 \int_s^K e^{-k_2 \mu_n \epsilon} d\epsilon + 4 \int_0^K \mu_n \beta_{a_n-d} d\epsilon \quad (75)$$

$$L_n(\Pi) \leq s + k'_1 \int_s^\infty e^{-k_2 \mu_n \epsilon} d\epsilon + 4 \int_0^K \mu_n \beta_{a_n-d} d\epsilon \quad (76)$$

$$= s + \frac{k'_1 e^{-k_2 \mu_n \epsilon}}{k_2 \mu_n} + k_3 \mu_n \beta_{a_n-d}. \quad (77)$$

Using [Assumption C](#), take  $a_n = n^{1/(1+\kappa)}$ ,  $\mu_n = n^{\kappa/(1+\kappa)}$ , and  $s = \frac{\log k'_1}{n^{\kappa/(1+\kappa)} k_2}$  to balance the exponential and linear terms. Then,

$$L_n(\Pi) = O\left(\sqrt{\frac{\text{VCD}(\mathcal{F}) \log n}{n^{\kappa/(1+\kappa)}}}\right). \quad (78)$$

For the lower bound, apply the IID version. ■

If we instead assume *algebraic mixing*, i.e.  $\beta_a = c_1 a^{-r}$ , then we can retrieve the same rate where  $0 < \kappa < (r - 1)/2$  (see Meir [37]). [Theorem 6.2](#) says that in dependent data settings, using the blocking approach developed here, we may pay a penalty: the upper bound on  $L_n(\Pi)$  goes to zero more slowly than in the IID case. But, the lower bound cannot be made any tighter since IID processes are still allowed under [Assumption C](#) (and of course under the more general [Assumption A](#)). In other words, we may have  $\kappa \rightarrow \infty$  so we can not rule out the faster learning rate of  $O\left(\sqrt{\frac{\text{VCD}(\mathcal{F}) \log n}{n}}\right)$ .

## 6.2 Structural risk minimization

Our presentation so far has focused on choosing one function  $\hat{f}$  from a model  $\mathcal{F}$  and demonstrating that the prediction risk  $R_n(\hat{f})$  is well characterized by the training error inflated by a complexity term. The procedure for actually choosing  $\hat{f}$  has been ignored. Common ways of choosing  $\hat{f}$  are frequently referred to as *empirical risk minimization* or ERM: approximate the expected risk  $R_n(f)$  with the empirical risk  $\hat{R}_n(f)$ , and choose  $\hat{f}$  to minimize the empirical risk. Many likelihood based methods have exactly this flavor. But more frequently, forecasters have many different models in mind, each with a different empirical risk minimizer. Regularized model classes (ridge regression, lasso, Bayesian methods) implicitly have this structure — altering the amount of regularization leads to different models  $\mathcal{F}$ . Or one may have many different forecasting models from which the forecaster would like to choose the best. This scenario leads to a generalization of ERM called *structural risk minimization* or SRM.

Given a collection of models  $\mathcal{F}_1, \mathcal{F}_2, \dots$  each with associated empirical risk minimizers  $\hat{f}_1, \hat{f}_2, \dots$ , we wish to use the function which has the smallest risk. Of course different models have different complexities, and those with larger complexities will tend to have smaller empirical risk. To choose the best function, we therefore penalize the empirical risk and select that function which minimizes the penalized version. Model selection tools like AIC or BIC have exactly this form, but they rely on specific knowledge of the data likelihood and use asymptotics to derive approximate penalties. In contrast, we have finite-sample bounds for the expected risk. This leads to a natural model selection rule: choose the predictor which has the smallest bound on the expected risk.

The generalization error bounds in [Section 4](#) allow one to perform model selection via the SRM principle without knowledge of the likelihood or appeals to asymptotic results. The penalty accounts for the complexity of the model through the VC dimension. Most useful however is that by using generalization error bounds for model selection, we are minimizing the prediction risk. So in the volatility forecasting exercise above, we would choose the mean.

If we want to make the prediction risk as small as possible, we can minimize the generalization error bound simultaneously over models  $\mathcal{F}$  and functions within those models. This amounts to treating VC dimension as a control variable. Therefore, by minimizing both the empirical risk and the VC dimension, we can choose that model and function which has the smallest prediction risk, a claim which other model selection procedures cannot make [\[32, 53\]](#).

## 7 Conclusion

This paper demonstrates how to control the generalization error of common macroeconomic forecasting models — ARMA models, vector autoregressions (Bayesian or otherwise), linearized dynamic stochastic general equilibrium models, and linear state-space models. We derive upper bounds on the risk, which hold with high probability while requiring only weak assumptions on the data-generating process. These bounds are finite sample in nature, unlike standard model selection penalties such as AIC or BIC. Furthermore, they do not suffer the biases inherent in other risk estimation techniques such as the pseudo-cross validation approach often used in the economic forecasting literature.

While we have stated these results in terms of standard economic forecasting models, they have very wide applicability. [Theorem 4.3](#) applies to any forecasting procedure with fixed memory length, linear or non-linear. [Theorem 4.6](#) applies only to methods whose forecasts are linear in the observations, but a similar result for nonlinear methods would just need to ensure that the dependence of the forecast on the past decays in some suitable way.

Rather than deriving bounds theoretically, one could attempt to estimate bounds on the risk. While cross-validation is tricky [44], nonparametric bootstrap procedures may do better. A fully nonparametric version is possible, using the circular bootstrap (reviewed in [31]). Bootstrapping lengthy out-of-sample sequences for testing fitted model predictions yields intuitively sensible estimates of  $R_n(f)$ , but there is currently no theory about the coverage level. Also, while models like VARs can be fit quickly to simulated data, general state-space models, let alone DSGEs, require large amounts of computational power, which is an obstacle to any resampling method.

While our results are a crucial first step for the learning-theoretic analysis of time series forecasts, many avenues remain for future exploration. To gain a more complete picture of the performance of forecasting algorithms, we would want minimax lower bounds (cf. [51]). These would tell us the smallest risk we could hope to achieve using any forecaster in some larger model class, letting us ask whether any of the models in common use actually approach this minimum. Another possibility is to target not the *ex ante* risk of the forecast, but the *ex post* regret: how much better might our forecasts have been, in retrospect and on the actually-realized data, had we used a different prediction function from the model  $\mathcal{F}$  [8, 45]? Remarkably, we can find forecasters which have low *ex post* regret, even if the data came from an adversary trying to make us perform badly. If we target regret rather than risk, we can actually ignore mixing, and even stationarity [48].

An increased recognition of the abilities and benefits of statistical learning theory can be of tremendous aid to financial and economic forecasters. The results presented here represent an initial yet productive foray in this direction. They allow for principled model comparisons as well as high probability performance guarantees. Future work in this direction will only serve to sharpen our ability to measure predictive power.

## A Auxiliary results

**Lemma A.1** (Lemma 4.1 in [56]). *Let  $Z$  be an event with respect to the block sequence  $\mathbf{U}$ . Then,*

$$|\mathbb{P}(Z) - \tilde{\mathbb{P}}(Z)| \leq \beta_a(\mu - 1), \quad (79)$$

where the first probability is with respect to the dependent block sequence,  $\mathbf{U}$ , and  $\tilde{\mathbb{P}}$  is with respect to the independent sequence,  $\mathbf{U}'$ .

This lemma essentially gives a method of applying IID results to  $\beta$ -mixing data. Because the dependence decays as we increase the separation between blocks, widely spaced blocks are nearly independent of each other. In particular, the difference between expectations over these nearly independent blocks and expectations over blocks which are actually independent can be controlled by the  $\beta$ -mixing coefficient.

**Lemma A.2** (Theorem 7 in Cortes et al. [10]). *Under [Assumption B](#),*

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \frac{R_n(f) - \hat{R}_n(f)}{Q_n(f)} > \epsilon \sqrt{2 + \log \frac{1}{\epsilon}}\right) \leq 4(2n + 1)^{\text{vcd}(\mathcal{F})} \exp\left\{-\frac{n\epsilon^2}{4}\right\} \quad (80)$$

**Corollary A.3.**

$$\begin{aligned} & \mathbb{P}\left(\sup_{f \in \mathcal{F}} \frac{R_n(f) - \hat{R}_n(f)}{Q_n(f)} > \epsilon\right) \\ & \leq 4(2n + 1)^{\text{vcd}(\mathcal{F})} \exp\left\{-\frac{n \exp\left(W\left(-\frac{2\epsilon^2}{e^4}\right) + 4\right)}{4}\right\}. \end{aligned} \quad (81)$$

## B Proofs of selected results

*Proof of [Theorem 4.5](#).* The VC dimension of a linear classifier  $f : \mathbb{R}^d \rightarrow \{0, 1\}$  is  $d$  (cf. Vapnik [53]). Real valued predictions have an extra degree of freedom.



For the VAR case, we are interested in the VC dimension of a multivariate linear classifier. Thus, one must be able to shatter collections of vectors where each vector is a binary sequence of length  $k$ . For a VAR, each coordinate is independent, thus, one can shatter a collection of vectors if one can shatter each coordinate projection. The result then follows from the AR case.  $\blacksquare$

*Proof of Theorem 4.6 and Corollary 4.7.* Let  $\mathcal{F}$  be indexed by the parameters of the growing memory model. Let  $\mathcal{F}'$  be the same class of models, but predictions are made based on the truncated memory length  $d$ . Define  $\tilde{R}_n(f')$  to be the training error of this truncated predictor  $f'$ . Then, for any  $f \in \mathcal{F}$ , and  $f' \in \mathcal{F}'$

$$R_n(f) - \hat{R}_n(f) \leq (R_n(f) - R_n(f')) + (R_n(f') - \tilde{R}_n(f')) + (\tilde{R}_n(f') - \hat{R}_n(f)). \quad (82)$$

We will need to handle all three terms. The first and third terms are similar. Let  $\mathbf{B}$  be as above and define the truncated linear predictor to have the same form but with  $\mathbf{B}$  replaced by

$$\mathbf{B}' = \begin{bmatrix} b_{d,1} & b_{d,2} & \cdots & b_{d,d} & & 0 \\ & b_{d+1,2} & \cdots & b_{d+1,d} & b_{d+1,d+1} & \\ & & & & \ddots & \\ 0 & & & & & b_{n,n-d+1} & \cdots & b_{n,n} \end{bmatrix}. \quad (83)$$

Then notice that

$$\tilde{R}_n(f') - \hat{R}_n(f) \leq |\tilde{R}_n(f') - \hat{R}_n(f)| \quad (84)$$

$$= \left| \frac{1}{n-d-1} \sum_{i=d}^{n-1} \ell(Y_{i+1} - \mathbf{b}_i Y_{i-d+1:i}) - \frac{1}{n-d-1} \sum_{i=d}^{n-1} \ell(Y_{i+1} - \mathbf{b}'_i Y_{i-d+1:i}) \right| \quad (85)$$

$$\leq \frac{\Delta}{n-d-1} \sum_{i=d}^{n-1} \ell((\mathbf{b}_i - \mathbf{b}'_i) Y_{i-d+1:i}) \quad (86)$$

by the triangle inequality where  $\mathbf{b}_i$  is the  $i^{\text{th}}$  row of  $\mathbf{B}$  and analogously for  $\mathbf{b}'_i$ . Therefore

$$\tilde{R}_n(f') - \hat{R}_n(f) \leq \frac{\Delta}{n-d-1} \sum_{i=d}^{n-1} \ell((\mathbf{b}_i - \mathbf{b}'_i) Y_{i-d+1:i}) \quad (87)$$

$$= \frac{\Delta}{n-d-1} \sum_{i=d}^{n-1} \ell \left( \sum_{j=1}^{i-d} b_{i,j} y_j \right) \quad (88)$$

For the case of the expected risk, we need only consider the first rows of  $\mathbf{B}$  and  $\mathbf{B}'$ . Using linearity of expectations and stationarity

$$R_n(f) - R_n(f') = |\mathbb{E}[\ell(Y_{n+1} - \mathbf{b}_n Y_{1:n+1})] - \mathbb{E}[\ell(Y_{n+1} - \mathbf{b}'_n Y_{1:n+1})]| \quad (89)$$

$$\leq \Delta \mathbb{E}[\ell((\mathbf{b}_n - \mathbf{b}'_n) Y_{1:n+1})] = \Delta \mathbb{E} \left[ \ell \left( \sum_{i=1}^{n-d} b_{n,i} Y_i \right) \right] \quad (90)$$

$$\leq \Delta^2 \sum_{i=1}^{n-d} \mathbb{E}[\ell(b_{n,i} Y_i)] \leq \Delta^2 \sum_{i=1}^{n-d} \ell(b_{n,i}) \mathbb{E}[\ell(Y_i)] \quad (91)$$

$$\leq \Delta^2 \mathbb{E}[\ell(Y_1)] \sum_{i=1}^{n-d} \ell(b_{n,i}) \quad (92)$$

Then,

$$R_n(f) - \hat{R}_n(f) - \delta_d(f) \leq R_n(f') - \tilde{R}_n(f') \quad (93)$$

where

$$\delta_d(f) = \Delta^2 \mathbb{E}[\ell(Y_1)] \sum_{j=1}^{n-d} \ell(b_{n,j}) + \frac{\Delta}{n-d-1} \sum_{i=d}^{n-1} \ell \left( \sum_{j=1}^{i-d} b_{i,j} y_j \right) \quad (94)$$

Divide through by  $Q_n(f)$  and take the supremum over  $\mathcal{F}$  and  $\mathcal{F}'$

$$\sup_{f \in \mathcal{F}} \frac{R_n(f) - \hat{R}_n(f) - \delta_d(f)}{Q_n(f)} \leq \sup_{f' \in \mathcal{F}', f \in \mathcal{F}} \frac{R_n(f') - \tilde{R}_n(f')}{Q_n(f)}. \quad (95)$$

Finally,

$$\sup_{f \in \mathcal{F}, f' \in \mathcal{F}'} \frac{Q_n(f')}{Q_n(f)} \leq 1 \quad (96)$$

since  $\mathcal{F}' \subseteq \mathcal{F}$ . So,

$$\sup_{f' \in \mathcal{F}', f \in \mathcal{F}} \frac{R_n(f') - \tilde{R}_n(f')}{Q_n(f)} = \sup_{f' \in \mathcal{F}', f \in \mathcal{F}} \frac{R_n(f') - \tilde{R}_n(f')}{Q_n(f')} \frac{Q_n(f')}{Q_n(f)} \quad (97)$$

$$\leq \sup_{f' \in \mathcal{F}'} \frac{R_n(f') - \tilde{R}_n(f')}{Q_n(f')}. \quad (98)$$

Now,

$$\mathbb{P} \left( \sup_{f \in \mathcal{F}} \frac{R_n(f) - \hat{R}_n(f) - \delta_d(f)}{Q_n(f)} > \epsilon \right) \leq \mathbb{P} \left( \sup_{f' \in \mathcal{F}'} \frac{R_n(f') - \tilde{R}_n(f')}{Q_n(f')} > \epsilon \right). \quad (99)$$

Since  $\mathcal{F}'$  is a class with finite memory, we can apply [Theorem 4.3](#) and [Corollary 4.4](#) to get the results. ■

*Proof of [Corollary 4.8](#).* This follows immediately from [Corollary 4.7](#) and (54). ■

## C Data

The data to estimate the RBC model is publicly available from the Federal Reserve Economic Database, FRED (<http://research.stlouisfed.org/fred2/>). The necessary series are shown in the [Table 2](#). All of the data is quarterly. The required series are PCESVC96, PCNDGC96, GDPIC1, HOANBS, and CNP16OV. These five series are used to create four series  $[y'_t, c'_t, i'_t, h'_t]$  as follows:

$$c'_t = 2.5 \times 10^5 \frac{PCESVC96 + PCNDGC96}{CNP16OV} \quad (100)$$

$$i'_t = 2.5 \times 10^5 \frac{GDPIC1}{CNP16OV} \quad (101)$$

$$y'_t = c_t + i_t \quad (102)$$

$$h'_t = 6000 \frac{HOANBS}{CNP16OV}. \quad (103)$$

We use the preprocessed data which accompanies DeJong and Dave [\[13\]](#) (<http://www.pitt.edu/~dejong/seconded.htm>). We then apply the HP-filter described in Hodrick and Prescott [\[25\]](#) to each series individually to calculate trend components  $[\tilde{y}_t, \tilde{c}_t, \tilde{i}_t, \tilde{h}_t]$ . The HP-filter amounts to fitting the smoothing spline

$$\tilde{\mathbf{x}}_{1:n} = \underset{\mathbf{z}_{1:n}}{\operatorname{argmin}} \sum_{t=1}^n (x'_t - z_t)^2 + \lambda \sum_{t=2}^{n-1} ((z_{t+1} - z_t) - (z_t - z_{t-1}))^2, \quad (104)$$

with the convention  $\lambda = 1600$ . We then calculate the detrended series that will be fed into the RBC model as

$$x_t = \log x'_t - \log \tilde{x}'_t. \quad (105)$$

The result is shown in [Figure 3](#).

Table 2: Data series from FRED			
Series ID	Description	Unit	Availability
PCESVC96	Real Personal Consumption Expenditures: Services	Billions of Chained 2005 \$	1/1/1995
PCNDGC96	Real Personal Consumption Expenditures: Nondurable Goods	Billions of Chained 2005 \$	1/1/1995
GDPIC1	Real Gross Domestic Investment	Billions of Chained 2005 \$	1/1/1947
HOANBS	Nonfarm Business Sector: Hours of All Persons	Index: 2005=100	1/1/1947
CNP16OV	Civilian Noninstitutional Population	Thousands of Persons	1/1/1948

Table 3: Priors, constraints, and parameter estimates for the RBC model.

Parameter	Estimate	Prior		Constraint	
		Mean	Variance	Lower	Upper
$\alpha$	0.24	0.29	$2.5 \times 10^{-2}$	0.1	0.5
$\beta$	0.99	0.99	$1.25 \times 10^{-3}$	0.90	1
$\phi$	4.03	1.5	2.5	1	5
$\varphi$	0.13	0.6	0.1	0	1
$\delta$	0.03 2	$2.5 \times 10^{-2}$	$1 \times 10^{-3}$	0	0.2
$\rho$	0.89	0.95	$2.5 \times 10^{-2}$	0.80	1
$\sigma_\epsilon$	$3.45 \times 10^{-5}$	$1 \times 10^{-4}$	$2 \times 10^{-5}$	0	0.05
$\sigma_y$	$1.02 \times 10^{-6}$	—	—	0	1
$\sigma_c$	$2.30 \times 10^{-5}$	—	—	0	1
$\sigma_i$	$6.11 \times 10^{-4}$	—	—	0	1
$\sigma_n$	$1.68 \times 10^{-4}$	—	—	0	1

## D Estimation

To estimate, we maximize the likelihood returned by the Kalman filter, penalized by priors on each of the “deep” parameters. This is because the likelihood surface is very rough and there exists some prior information about the parameters. Additionally, each of the parameters is constrained to lie in a plausible interval. Each parameter has a normal prior with means and variances similar to those in the literature. We generally follow those in DeJong et al. [14]. The priors, constraints (which are strict), and estimates are shown in Table 3.

## References

- [1] ADAMS, T., AND NOBEL, A. (2010), “Uniform convergence of Vapnik-Chervonenkis classes under ergodic sampling,” *The Annals of Probability*, **38**(4), 1345–1367.
- [2] ATHANASOPOULOS, G., AND VAHID, F. (2008), “VARMA versus VAR for macroeconomic forecasting,” *Journal of Business and Economic Statistics*, **26**(2), 237–252.

- [3] BARTLETT, P. L., AND MENDELSON, S. (2002), “Rademacher and Gaussian complexities: Risk bounds and structural results,” *Journal of Machine Learning Research*, **3**, 463–482.
- [4] BOUSQUET, O., AND ELISSEEFF, A. (2001), “Algorithmic stability and generalization performance,” in *Advances in Neural Information Processing Systems*, vol. 13, pp. 196–202, Cambridge, MA, MIT Press.
- [5] BOUSQUET, O., AND ELISSEEFF, A. (2002), “Stability and generalization,” *The Journal of Machine Learning Research*, **2**, 499–526.
- [6] BRADLEY, R. C. (2005), “Basic properties of strong mixing conditions. A survey and some open questions,” *Probability Surveys*, **2**, 107–144.
- [7] CARRASCO, M., AND CHEN, X. (2002), “Mixing and moment properties of various GARCH and stochastic volatility models,” *Econometric Theory*, **18**(01), 17–39.
- [8] CESA-BIANCHI, N., AND LUGOSI, G. (2006), *Prediction, learning, and games*, Cambridge Univ Press, Cambridge, UK.
- [9] CORLESS, R., GONNET, G., HARE, D., JEFFREY, D., AND KNUTH, D. (1996), “On the Lambert  $w$  function,” *Advances in Computational Mathematics*, **5**(1), 329–359.
- [10] CORTES, C., MANSOUR, Y., AND MOHRI, M. (2010), “Learning bounds for importance weighting,” in *Advances in Neural Information Processing Systems 23*, eds. J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, vol. 23, pp. 442–450, MIT Press.
- [11] CRISTIANINI, N., AND SHAWE-TAYLOR, J. (2000), *An introduction to support Vector Machines and other kernel-based learning methods*, Cambridge Univ Press, Cambridge, UK.
- [12] DASGUPTA, A. (2008), *Asymptotic theory of statistics and probability*, Springer Texts in Statistics, Springer Verlag, New York.
- [13] DEJONG, D., AND DAVE, C. (2011), *Structural macroeconometrics*, Princeton Univ Press, Princeton, 2 edn.
- [14] DEJONG, D. N., INGRAM, B. F., AND WHITEMAN, C. H. (2000), “A Bayesian approach to dynamic macroeconomics,” *Journal of Econometrics*, **98**(2), 203–223.
- [15] DEJONG, D. N., DHARMARAJAN, H., LIESENFELD, R., MOURA, G. V., AND RICHARD, J.-F. (2009), “Efficient likelihood evaluation of state-space representations,” Tech. rep., University of Pittsburgh.
- [16] DEL NEGRO, M., SCHORFHEIDE, F., SMETS, F., AND WOUTERS, R. (2007), “On the fit and forecasting performance of New Keynesian models,” *Journal of Business and Economic Statistics*, **25**(2), 123–162.
- [17] DOUKHAN, P. (1994), *Mixing: Properties and Examples*, Springer Verlag, New York.
- [18] DURBIN, J., AND KOOPMAN, S. (2001), *Time Series Analysis by State Space Methods*, Oxford Univ Press, Oxford.
- [19] FAUST, J., AND WRIGHT, J. H. (2009), “Comparing Greenbook and reduced form forecasts using a large realtime dataset,” *Journal of Business and Economic Statistics*, **27**(4), 468–479.
- [20] FERNÁNDEZ-VILLAYERDE, J. (2009), “The econometrics of DSGE models,” Tech. rep., NBER Working Paper Series.
- [21] GERALI, A., NERI, S., SESSA, L., AND SIGNORETTI, F. (2010), “Credit and banking in a DSGE model of the Euro area,” *Journal of Money, Credit and Banking*, **42**, 107–141.
- [22] GERTLER, M., AND KARADI, P. (2011), “A model of unconventional monetary policy,” *Journal of Monetary Economics*, **58**, 17–34.

- [23] GOODHART, C., OSORIO, C., AND TSOMOCOS, D. (2009), “Analysis of monetary policy and financial stability: A new paradigm,” Tech. Rep. 2885, CESifo.
- [24] HARVEY, A., RUIZ, E., AND SHEPHARD, N. (1994), “Multivariate stochastic variance models,” *The Review of Economic Studies*, **61**(2), 247–264.
- [25] HODRICK, R. J., AND PRESCOTT, E. C. (1997), “Postwar U.S. business cycles: An empirical investigation,” *Journal of Money, Credit, and Banking*, **29**(1), 1–16.
- [26] HOEFFDING, W. (1963), “Probability inequalities for sums of bounded random variables,” *Journal of the American Statistical Association*, **58**(301), 13–30.
- [27] KALMAN, R. E. (1960), “A new approach to linear filtering and prediction problems,” *Journal of Basic Engineering*, **82**(1), 35–45.
- [28] KEARNS, M., AND RON, D. (1999), “Algorithmic stability and sanity-check bounds for leave-one-out cross-validation,” *Neural Computation*, **11**(6), 1427–1453.
- [29] KLEIJN, B. J. K., AND VAN DER VAART, A. W. (2006), “Misspecification in infinite-dimensional Bayesian statistics,” *Annals of Statistics*, **34**, 837–877.
- [30] KYDLAND, F. E., AND PRESCOTT, E. C. (1982), “Time to build and aggregate fluctuations,” *Econometrica*, **50**(6), 1345–1370.
- [31] LAHIRI, S. N. (1999), “Theoretical comparisons of block bootstrap methods,” *Annals of Statistics*, **27**(1), 386–404.
- [32] MASSART, P. (2007), “Concentration inequalities and model selection,” in *Ecole d’Eté de Probabilités de Saint-Flour XXXIII-2003*, Springer.
- [33] MAYO, D. G. (1996), *Error and the Growth of Experimental Knowledge*, University of Chicago Press, Chicago.
- [34] McDONALD, D. J., SHALIZI, C. R., AND SCHERVISH, M. (2011a), “Estimating  $\beta$ -mixing coefficients,” in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, eds. G. Gordon, D. Dunson, and M. Dudík, vol. 15, JMLR W&CP.
- [35] McDONALD, D. J., SHALIZI, C. R., AND SCHERVISH, M. (2011b), “Estimating  $\beta$ -mixing coefficients via histograms,” submitted for publication.
- [36] McDONALD, D. J., SHALIZI, C. R., AND SCHERVISH, M. (2011c), “Generalization error bounds for stationary autoregressive models,” .
- [37] MEIR, R. (2000), “Nonparametric time series prediction through adaptive model selection,” *Machine Learning*, **39**(1), 5–34.
- [38] MOHRI, M., AND ROSTAMIZADEH, A. (2009), “Rademacher complexity bounds for non-iid processes,” in *Advances in Neural Information Processing Systems*, eds. D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, vol. 21, pp. 1097–1104, MIT Press, Cambridge, MA.
- [39] MOHRI, M., AND ROSTAMIZADEH, A. (2010), “Stability bounds for stationary  $\varphi$ -mixing and  $\beta$ -mixing processes,” *Journal of Machine Learning Research*, **11**, 789–814.
- [40] MOKKADEM, A. (1988), “Mixing properties of ARMA processes,” *Stochastic Processes and their Applications*, **29**(2), 309–315.
- [41] MÜLLER, U. K. (2011), “Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix,” Tech. rep., Princeton University.
- [42] POLLARD, D. (1984), *Convergence of stochastic processes*, Springer Verlag, New York.

- [43] POLLARD, D. (1990), *Empirical processes: Theory and applications*, Institute of Mathematical Statistics.
- [44] RACINE, J. (2000), “Consistent cross-validators model-selection for dependent data: HV-block cross-validation,” *Journal of Econometrics*, **99**(1), 39–61.
- [45] RAKHLIN, A., SRIDHARAN, K., AND TEWARI, A. (2010), “Online learning: Random averages, combinatorial parameters, and learnability,” .
- [46] ROMER, D. (2011), *Advanced macroeconomics*, McGraw-Hill, 4 edn.
- [47] SHALIZI, C. R. (2009), “Dynamics of Bayesian updating with dependent data and misspecified models,” *Electronic Journal of Statistics*, **3**, 1039–1074.
- [48] SHALIZI, C. R., JACOBS, A. Z., KLINKNER, K. L., AND CLAUSET, A. (2011), “Adapting to non-stationarity with growing expert ensembles,” .
- [49] SIMS, C. A. (2002), “Solving linear rational expectations models,” *Computational Economics*, **20**(1-2), 1–20.
- [50] SMETS, F., AND WOUTERS, R. (2007), “Shocks and frictions in US business cycles: A Bayesian DSGE approach,” *American Economic Review*, **97**(3), 586–606.
- [51] TSYBAKOV, A. (2009), *Introduction to nonparametric estimation*, Springer Verlag.
- [52] VAPNIK, V. (1998), *Statistical learning theory*, John Wiley & Sons, Inc., New York.
- [53] VAPNIK, V. (2000), *The Nature of Statistical Learning Theory*, Springer Verlag, New York, 2nd edn.
- [54] VAPNIK, V., AND CHERVONENKIS, A. (1971), “On the uniform convergence of relative frequencies of events to their probabilities,” *Theory of Probability and its Applications*, **16**, 264–280.
- [55] WASSERMAN, L. (2006), *All of Nonparametric Statistics: A Concise Course in Nonparametric Statistical Inference*, Springer Verlag.
- [56] YU, B. (1994), “Rates of convergence for empirical processes of stationary mixing sequences,” *The Annals of Probability*, **22**(1), 94–116.